1–8

A New Analysis of Compressive Sensing by Stochastic Proximal Gradient Descent

Rong Jin[†] RONGJIN@CSE.MSU.EDU
Tianbao Yang[‡] TYANG@GE.COM
Shenghuo Zhu[♮] ZSH@SV.NEC-LABS.COM
[†]*Department of Computer Science and Engineering*
Michigan State University, East Lansing, MI, 48824, USA
[‡]*Machine Learning Lab, GE Global Research, San Ramon, CA, 94583, USA*
[♮]*NEC Laboratories America, Cupertino, CA, 95014, USA*

Abstract

In this manuscript, we analyze the sparse signal recovery (compressive sensing) problem from the perspective of convex optimization by stochastic proximal gradient descent. This view allows us to significantly simplify the recovery analysis of compressive sensing. More importantly, it leads to an efficient optimization algorithm for solving the regularized optimization problem related to the sparse recovery problem. Compared to the existing approaches, there are two advantages of the proposed algorithm. First, it enjoys a geometric convergence rate and therefore is computationally efficient. Second, it guarantees that the support set of any intermediate solution generated by the proposed algorithm is concentrated on the support set of the optimal solution.

1. Introduction and Related Work

The problem of sparse signal recovery is to reconstruct a sparse signal given a number of linear measurements of the signal. The problem has been studied extensively under two closely related settings, i.e., lasso and compressive sensing . Lasso is known as a tool of model selection that aims to learn a sparse model $\beta \in \mathbb{R}^d$ from a data design matrix $X \in \mathbb{R}^{n\times d}$ and noisy measurements $\mathbf{y} = X\beta + \varepsilon$ of β , where ε are zero-mean independent Gaussian random variables, by solving the ℓ_1 regularized least square problem $\min_{\beta\in\mathbb{R}^d} \|\mathbf{y} - X\beta\|_2^2 + \lambda\|\beta\|_1$. Compressive sensing focuses more on the study of how many random measurements are needed to optimally recover a sparse signal $\mathbf{x}_* \in \mathbb{R}^d$. In the manuscript, we provide a new perspective of compressive sensing from the viewpoint of convex optimization by gradient descent. Our analysis reveals that in order to solve the optimal recovery problem of $\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{x} - \mathbf{x}_*\|_2^2$ in hindsight by a gradient descent method, the random measurements of the signal $\mathbf{x}_*$ denoted by $U\mathbf{x}_*$ are used for computing a stochastic gradient of the objective. Furthermore, we develop a stochastic gradient descent method that solves a composite gradient mapping with ℓ_1 regularization at each iteration, which ensures the support set of intermediate solution concentrates on the support set of the optimal solution. Finally, we prove that the proposed algorithm enjoys a geometric convergence rate. To the best of our knowledge, this work is the first that analyze the compressive sensing in the angle of optimization by stochastic gradient descent.

A great volume of work have been devoted to the problem of sparse signal recovery in different philosophies. In the following, we briefly review some related work that solves the optimization problem for reconstructing the optimal signal with a linear (i.e., geometric) convergence rate. In (Bredies and Lorenz, 2008; Hale et al., 2008), the authors established linear convergence rates as the iterates are close enough to the optimum. Tropp and Gilbert (2007) showed that if an algorithm can quickly identify the support set of the optimal solution, then the optimization is effectively reduced to a lower-dimensional subspace, and geometric convergence can be achieved. Garg and Khandekar (2009) showed a geometric convergence rate for the recovered solution by a sparsification. In (Agarwal et al., 2011), the authors showed that a simple gradient descent algorithm for the constrained Lasso can achieve a global geometric convergence rate in recovering the target solution (Corollary 2) [1]. One shortcoming with the analysis in (Agarwal et al., 2011) is that the parameter κ in linear convergence is lower bounded by a constant (i.e., 3/4) independent from the number of random measurements, a disappointing feature as we expect a faster convergence with the increasing number of random measurements.

The proposed approach is similar to several existing algorithms (Wen et al., 2010; Wright et al., 2009; Hale et al., 2008; Xiao and Zhang, 2012) developed for ℓ_1 regularized minimization in that all of them solve the regularized optimization problem by gradually shrinking the value of the regularization parameter. To the best of our knowledge, (Xiao and Zhang, 2012) is the only work in this direction that provides theoretical guarantee. The main difference between this work and the work (Xiao and Zhang, 2012) is that instead of performing a simple gradient mapping for each value of the regularized parameter, the algorithm (Xiao and Zhang, 2012) requires, at each iteration, solving the $L1$ regularized optimization problem to certain accuracy, leading to a significant computational overhead in optimization.

2. Algorithm

Let $\mathbf{x}_* \in \mathbb{R}^d$ be a s-sparse high dimensional signal to be recovered, where the number of non-zero elements in $\mathbf{x}_*$ is s. We denote by $S(\mathbf{x})$ the support set for $\mathbf{x}$ that includes all the indices of the non-zero entries in $\mathbf{x}$, i.e.,

$$S(\mathbf{x}) = \{i \in [d] : [\mathbf{x}]_i \neq 0\} \tag{1}$$

where $[d]$ denotes the set $\{1, \ldots, d\}$ and $[\mathbf{x}]_i$ denote the i-th element in $\mathbf{x}$. We also denote by $\overline{\mathcal{S}}(\mathbf{x}) = [d] \setminus \mathcal{S}(\mathbf{x})$ the complementary set of $\mathcal{S}(\mathbf{x})$. In particular, we use $\mathcal{S}_*, \overline{\mathcal{S}}_*$ to denote the support set and complementary set of $\mathbf{x}_*$. Similar to most of the previous analysis, we assume that $\|\mathbf{x}_*\|_2 \leq R$.

To motivate our approach, we first consider the following optimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^d} \quad \mathcal{L}(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{x}_*\|_2^2 \tag{2}$$

1. In the same paper, the authors also discussed a gradient descent algorithm for the regularized Lasso, which unfortunately is only able to recover the solution up to the statistical tolerance.

Evidently, the optimal solution to (2) is $\mathbf{x}_*$. We now consider a gradient descent method for optimizing the problem in (2), leading to the following updating equation for $\mathbf{x}_t$

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{x} - (\mathbf{x}_t - \nabla\mathcal{L}(\mathbf{x}_t))\|_2^2 \tag{3}$$

where $\nabla\mathcal{L}(\mathbf{x}) = \mathbf{x} - \mathbf{x}_*$. Since the problem in (2) is both smooth and strongly convex, the above updating enjoys a geometric convergence rate [2], allowing an efficient reconstruction of $\mathbf{x}_*$.

However, the updating rule in (3) can not be used because it requires knowing $\mathbf{x}_*$, the full information of the sparse signal to be recovered. In compressive sensing, the only available information about the target signal $\mathbf{x}_*$ is its random measurements. More specifically, let $U \in \mathbb{R}^{m\times d}$ be a random measurement matrix and $\mathbf{y} = U\mathbf{x}_*$ be the corresponding m random measurements. Using the random measurements, we construct an approximate gradient as

$$\widehat{\nabla}\mathcal{L}(\mathbf{x}_t) = U^\top U(\mathbf{x}_t - \mathbf{x}_*) = U^\top(U\mathbf{x}_t - \mathbf{y}) \tag{4}$$

To ensure $\widehat{\nabla}\mathcal{L}(\mathbf{x}_t)$ provide an useful estimate of $\nabla\mathcal{L}(\mathbf{x}_t)$, we assume the random measurement matrix U satisfies the following restricted isometry properties (RIP) (with an overwhelming probability).

Definition 1 (s-restricted isometry constant) *Let $\delta_s \geq 0$ be the smallest constant such that for any subset $\mathcal{T} \in [d]$ with $|\mathcal{T}| \leq s$ and $\mathbf{x} \in \mathbb{R}^{|\mathcal{T}|}$,*

$$(1-\delta_s)\|\mathbf{x}\|_2^2 \leq \|U_\mathcal{T}\mathbf{x}\|_2^2 \leq (1+\delta_s)\|\mathbf{x}\|_2^2$$

where $U_\mathcal{T}$ denote the sub-matrix of U with columns from $\mathcal{T}$.

Definition 2 (s,s-restricted orthogonality constant) *Let $\theta_{s,s}$ be the smallest constant such that for any two disjoint subsets $\mathcal{T},\mathcal{T}' \in [d]$ with $|\mathcal{T}| \leq s$, $|\mathcal{T}'| \leq s$, $2s \leq d$, and for any $\mathbf{x} \in \mathbb{R}^{|\mathcal{T}|}$, $\mathbf{x}' \in \mathbb{R}^{|\mathcal{T}'|}$,*

$$|\langle U_\mathcal{T}\mathbf{x}, U_{\mathcal{T}'}\mathbf{x}'\rangle| \leq \theta_{s,s}\|\mathbf{x}\|_2\|\mathbf{x}'\|_2$$

The above two constants are standard tools in the analysis of optimal recovery of compressive sensing. It has been shown that several random measurement matrix including Gaussian measurement matrix, binary measurement matrix, Fourier measurement matrix and incoherent measurement matrix satisfy the above RIP with small δ_s and $\theta_{s,s}$.

Next, we will use $\widehat{\nabla}\mathcal{L}(\mathbf{x}_t)$ as an approximation of $\nabla\mathcal{L}(\mathbf{x}_t)$ and update the solution by performing the following proximal mapping:

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \tau_t\|\mathbf{x}\|_1 + \langle \mathbf{x} - \mathbf{x}_t, \widehat{\nabla}\mathcal{L}(\mathbf{x}_t)\rangle + \frac{1+\gamma}{2}\|\mathbf{x} - \mathbf{x}_t\|_2^2 \tag{5}$$

where $\tau_t > 0$ is the regularization parameter that varies over the iterations and $\gamma > 0$ is a parameter essentially due to the RIP conditions. The updating rule given in (6) differs from (3) in that (i) the true gradient $\nabla\mathcal{L}(\mathbf{x}_t)$ is replaced with an approximate gradient $\widehat{\nabla}\mathcal{L}(\mathbf{x}_t)$ and (ii) a ℓ_1 regularization term $\tau_t\|\mathbf{x}\|_1$ is added. With appropriate choice of τ_t, this regularization term will essentially remove the noise arising from the approximate gradient and consequentially lead to the geometric convergence rate.

2. In fact, only one step is needed.

Algorithm 1 A Composite Optimization Approach for Compressive Sensing

1: **Input:** Gaussian random matrix $U \in \mathbb{R}^{d\times m}$, random measurements $\mathbf{y} = U^\top \mathbf{x}_*$, regularization parameters $\tau_1, \ldots, \tau_T$, and γ
2: **Initialize** $\mathbf{x}_1 = 0$.
3: **for** $t = 1, \ldots, T$ **do**
4: Compute $\widehat{\mathbf{x}}_t = \mathbf{x}_t - \frac{1}{1+\gamma} U(U^\top \mathbf{x}_t - \mathbf{y})$
5: Update the solution $\mathbf{x}_{t+1} = sign(\widehat{\mathbf{x}}_t)\left[|\widehat{\mathbf{x}}_t| - \frac{\tau_t}{1+\gamma}\right]_+$
6: **end for**
7: **Output** the final solution $\mathbf{x}_{T+1}$

Remark: We note that our approach is fundamentally different from the classical idea of stochastic gradient descent. In stochastic gradient descent, we have access to the stochastic oracle of the gradients. By drawing an unbiased estimate of the gradient independently from the statistical oracle at each iteration, stochastic gradient descent is able to reduce the noise in the stochastic gradients through the average by exploring the concentration inequality of martingales. In contrast, in compressive sensing, we are only provided with *one* set of random measurements for the target signal $\mathbf{x}_*$. Since all the estimates of gradients are based on the same set of random measurements, they are statistically dependent, making it impossible to explore the martingale technique for reducing the noise in the estimates of gradients. The ℓ_1 regularization term in the updating rule in (5) is essentially introduced to reduce the noise in the statistical gradients, and therefore plays similar role as the concentration inequality of martingales.

To give the solution of $\mathbf{x}_{t+1}$ in a closed form, we write (5) as

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\left\|\mathbf{x} - \left(\mathbf{x}_t - \frac{1}{1+\gamma}\widehat{\nabla}\mathcal{L}(\mathbf{x}_t)\right)\right\|_2^2 + \frac{\tau_t}{1+\gamma}\|\mathbf{x}\|_1 \tag{6}$$

According to , the value of $\mathbf{x}_{t+1}$ is given by

$$\mathbf{x}_{t+1} = sign(\widehat{\mathbf{x}}_t)\left[|\widehat{\mathbf{x}}_t| - \frac{\tau_t}{1+\gamma}\right]_+ \tag{7}$$

where $\widehat{\mathbf{x}}_t = \mathbf{x}_t - (1/(1+\gamma))\widehat{\nabla}\mathcal{L}(\mathbf{x}_t)$ and $[v]_+ = \max(0, v)$. We present the detailed steps of the proposed approach in Algorithm 1 for reconstructing the sparse signal given a set of random measurements. To end this section, we present our main result in the following theomrem which states the theoretical guarantee of Algorithm 1.

Theorem 1 *Let $\mathbf{x}_* \in \mathbb{R}^d$ be a s-sparse signal and $\mathbf{y} = U\mathbf{x}_*$ be a set of m random measurements of $\mathbf{x}_*$. Set γ, τ_t in Algorithm 1 as*

$$\gamma = \max(\delta_{3s}, \theta_{s,s} + \delta_s), \quad \tau_t = \frac{\theta_{s,s} + \delta_s + \gamma}{\sqrt{s}}(4\gamma)^{(t-1)/2}R, t = 1, \ldots, T.$$

If we assume $\gamma \le 1/4$, then (i) $\|\mathcal{S}_t \cup \mathcal{S}_*\| \le 2s$ and (ii) $\|\mathbf{x}_t - \mathbf{x}_*\|_2 \le (4\gamma)^{(t-1)/2}\|\mathbf{x}_*\|_2$, and (iii) $\|\mathbf{x}_t - \mathbf{x}_*\|_1 \le \sqrt{s}(4\gamma)^{(t-1)/2}\|\mathbf{x}_*\|_2, t = 1, \ldots, T$

3. Analysis

Before presenting our analysis, we introduce a few notations that will be used throughout the paper. Given a set $\mathcal{S} \subseteq [d]$, we denote $[\mathbf{x}]_{\mathcal{S}}$ the vector that only includes the entries of $\mathbf{x}$ in the subset $\mathcal{S}$. Given two subsets $\mathcal{A} \subseteq [d]$ and $\mathcal{B} \subseteq [d]$, we denote by $[M]_{\mathcal{A},\mathcal{B}}$ a sub-matrix that includes all the entries (i,j) in matrix M with $i \in \mathcal{A}$ and $j \in \mathcal{B}$. We first prove the following Theorem.

Theorem 2 *Let $\mathcal{S}_t$ be the support set of $\mathbf{x}_t$ and $\mathcal{S}_*$ be the support set of $\mathbf{x}_*$. Define $\mathcal{S}_t^c = \mathcal{S}_t \cup \mathcal{S}_*$, $\mathcal{S}_t^a = \mathcal{S}_c \backslash \mathcal{S}_*$. If we assume $|\mathcal{S}_t \cup \mathcal{S}_*| \leq 2s$, at most s entries of $[(1+\gamma)\mathbf{x}_t - U^\top U(\mathbf{x}_t - \mathbf{x}_*)]_{\overline{\mathcal{S}}_*}$ with magnitude larger than $\dfrac{\theta_{s,s} + \theta_{s,s} + \gamma}{\sqrt{s}} \|\mathbf{x}_t - \mathbf{x}_*\|_2$..*

Proof For any subset $\mathcal{S}' \subset \overline{\mathcal{S}}_*$ of size s, let $\mathcal{S}_1' = \mathcal{S}' \cap \mathcal{S}_t^a$ and $\mathcal{S}_2' = \mathcal{S}' \setminus \mathcal{S}_t^a$. We have

$$
\begin{aligned}
&\left\|[U^\top U(\mathbf{x}_t - \mathbf{x}_*)]_{\mathcal{S}'} - (1+\gamma)[\mathbf{x}_t]_{\mathcal{S}'}\right\|_2 = \left\|U_{\mathcal{S}'}^\top U_{\mathcal{S}_*}[\mathbf{x}_t - \mathbf{x}_*]_{\mathcal{S}_*} + U_{\mathcal{S}'}^\top U_{\mathcal{S}_t^a}[\mathbf{x}_t]_{\mathcal{S}_t^a} - (1+\gamma)[\mathbf{x}_t]_{\mathcal{S}'}\right\|_2 \\
&\leq \left\|U_{\mathcal{S}'}^\top U_{\mathcal{S}_*}[\mathbf{x}_t - \mathbf{x}_*]_{\mathcal{S}_*}\right\|_2 + \left\|U_{\mathcal{S}_2'}^\top U_{\mathcal{S}_t^a}[\mathbf{x}_t]_{\mathcal{S}_t^a}\right\|_2 + \left\|U_{\mathcal{S}_1'}^\top U_{\mathcal{S}_t^a}[\mathbf{x}_t]_{\mathcal{S}_t^a} - (1+\gamma)[\mathbf{x}_t]_{\mathcal{S}_1'}\right\|_2 \\
&\leq \left\|U_{\mathcal{S}'}^\top U_{\mathcal{S}_*}\right\|_2 \left\|[\mathbf{x}_t - \mathbf{x}_*]_{\mathcal{S}_*}\right\|_2 + \left\|U_{\mathcal{S}_2'}^\top U_{\mathcal{S}_t^a}\right\|_2 \left\|[\mathbf{x}_t]_{\mathcal{S}_t^a}\right\|_2 + \left\|U_{\mathcal{S}_t^a}^\top U_{\mathcal{S}_t^a}[\mathbf{x}_t]_{\mathcal{S}_t^a} - (1+\gamma)[\mathbf{x}_t]_{\mathcal{S}_t^a}\right\|_2 \\
&\leq \theta_{s,s}\|[\mathbf{x}_t - \mathbf{x}_*]_{\mathcal{S}_*}\|_2 + \theta_{s,s}\|[\mathbf{x}_t]_{\mathcal{S}_t^a}\|_2 + (\delta_s + \gamma)\|[\mathbf{x}_t]_{\mathcal{S}_t^a}\|_2 \leq (\theta_{s,s} + \delta_s + \gamma)\|\mathbf{x}_t - \mathbf{x}_*\|_2
\end{aligned}
$$

Since the above inequality holds for any subset $\mathcal{S}' \subseteq \overline{\mathcal{S}}_*$ of size s, we form the set $\mathcal{S}'$ by including the largest s entries in absolute value of $[(1+\gamma)\mathbf{x}_t - U^\top U(\mathbf{x}_t - \mathbf{x}_*)]_{\overline{\mathcal{S}}_*}$. Then the smallest absolute value in $[(1+\gamma)\mathbf{x}_t - U^\top U(\mathbf{x}_a - \mathbf{x}_b)]_{\mathcal{S}'}$ is bounded by $\dfrac{\theta_{s,s_a} + \theta_{s,s_a^c}}{\sqrt{s}}$. By the construction of $\mathcal{S}'$, the smallest entry in $\mathcal{S}'$ is the sth largest entry in $[(1+\gamma)\mathbf{x}_t - U^\top U(\mathbf{x}_t - \mathbf{x}_*)]_{\overline{\mathcal{S}}_*}$, we conclude that at most s entries with magnitude larger than $\dfrac{\theta_{s,s} + +\delta_s + \gamma}{\sqrt{s}}\|\mathbf{x}_t - \mathbf{x}_*\|_2$. ■

As an immediate result of Theorem 2, we prove the following Corollary.

Corollary 3 *Let S_t be the support set of $\mathbf{x}_t$ and $\mathcal{S}_*$ be the support set of $\mathbf{x}_*$. If $|S_t \cup S_*| \leq 2s$ and $\tau_t \geq \frac{\theta_{s,s}+\delta_s+\gamma}{\sqrt{s}}\|\mathbf{x}_t - \mathbf{x}_*\|_2$, then $|\mathcal{S}_{t+1} \cup \mathcal{S}_*| \leq 2s$ and $|\mathcal{S}_* \cup \mathcal{S}_t \cup \mathcal{S}_{t+1}| \leq 3s$.*

Proof As shown in (7), $\mathbf{x}_{t+1}$ is given by

$$
\mathbf{x}_{t+1} = sign(\widehat{\mathbf{x}}_t)\frac{1}{1+\gamma}\left[\left|(1+\gamma)\mathbf{x}_t - \widehat{\nabla}\mathcal{L}(\mathbf{x}_t)\right| - \tau_t\right]_+
$$

By Theorem 2, we know that there are at most s entries in $\left|\left[(1+\gamma)\mathbf{x}_t - \widehat{\nabla}\mathcal{L}(\mathbf{x}_t)\right]_{\overline{\mathcal{S}}_*}\right|$ are larger than $(\theta_{s,s} + \delta_s + \gamma)\|\mathbf{x}_t - \mathbf{x}_*\|_2/\sqrt{s}$, therefore $[\mathbf{x}_{t+1}]_{\overline{\mathcal{S}}_*}$ has at most s non-zeros entries. It concludes that $|\mathcal{S}_{t+1} \cup \mathcal{S}_*| \leq 2s$ and $|\mathcal{S}_* \cup \mathcal{S}_t \cup \mathcal{S}_{t+1}| \leq 3s$. ■

Theorem 4 *If we assume* $\|\mathbf{x}_t - \mathbf{x}_*\|_2^2 \leq \Delta_t^2 = (4\gamma)^{t-1}R^2$, *set* $\tau_t = \frac{\theta_{s,s}+\delta_s+\gamma}{\sqrt{s}}\Delta_t$ *and* $\gamma \geq \max(\delta_{3s}, \theta_{s,s}+\delta_s)$, *then we have*

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2^2 \leq \Delta_{t+1}^2 = (4\gamma)^t R^2$$

Proof Let $\mathcal{T} = \mathcal{S}_* \cup \mathcal{S}_t \cup \mathcal{S}_{t+1}$, by Corollary 3, we have $|\mathcal{T}| \leq 3s$, therefore $\|U_{\mathcal{T}}^\top U_{\mathcal{T}} - I\|_2 \leq \delta_{3s}$. Next, we proceed the proof as follows:

$$\begin{aligned}
\mathcal{L}(\mathbf{x}_{t+1}) &= \mathcal{L}(\mathbf{x}_t) + \langle \mathbf{x}_{t+1} - \mathbf{x}_t, \nabla\mathcal{L}(\mathbf{x}_t)\rangle + \frac{1}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 \\
&= \mathcal{L}(\mathbf{x}_t) + \langle \mathbf{x}_{t+1} - \mathbf{x}_t, \widehat{\nabla}\mathcal{L}(\mathbf{x}_t)\rangle + \langle \mathbf{x}_{t+1} - \mathbf{x}_t, \nabla\mathcal{L}(\mathbf{x}_t) - \widehat{\nabla}\mathcal{L}(\mathbf{x}_t)\rangle + \frac{1}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 \\
&\leq \mathcal{L}(\mathbf{x}_t) + \langle \mathbf{x}_{t+1} - \mathbf{x}_t, \widehat{\nabla}\mathcal{L}(\mathbf{x}_t)\rangle + \langle \mathbf{x}_{t+1} - \mathbf{x}_t, (I - U^\top U)(\mathbf{x}_t - \mathbf{x}_*)\rangle + \frac{1}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 \\
&\leq \mathcal{L}(\mathbf{x}_t) + \langle \mathbf{x}_{t+1} - \mathbf{x}_t, \widehat{\nabla}\mathcal{L}(\mathbf{x}_t)\rangle + \|I - U_{\mathcal{T}}^\top U_{\mathcal{T}}\|_2\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2\|\mathbf{x}_t - \mathbf{x}_*\|_2 + \frac{1}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 \\
&\leq \mathcal{L}(\mathbf{x}_t) + \langle \mathbf{x}_{t+1} - \mathbf{x}_t, \widehat{\nabla}\mathcal{L}(\mathbf{x}_t)\rangle + \delta_{3s}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2\|\mathbf{x}_t - \mathbf{x}_*\|_2 + \frac{1}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 \\
&\leq \mathcal{L}(\mathbf{x}_t) + \langle \mathbf{x}_{t+1} - \mathbf{x}_t, \widehat{\nabla}\mathcal{L}(\mathbf{x}_t)\rangle + \frac{1+\gamma}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 + \frac{\delta_{3s}^2}{2\gamma}\|\mathbf{x}_t - \mathbf{x}_*\|_2^2 \\
&\leq \mathcal{L}(\mathbf{x}_t) + \langle \mathbf{x}_{t+1} - \mathbf{x}_t, \widehat{\nabla}\mathcal{L}(\mathbf{x}_t)\rangle + \tau_t\|\mathbf{x}_{t+1}\|_1 + \frac{1+\gamma}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 \\
&+ \frac{\delta_{3s}^2}{2\gamma}\|\mathbf{x}_t - \mathbf{x}_*\|_2^2 - \tau_t\|\mathbf{x}_{t+1}\|_1 \\
&\leq \mathcal{L}(\mathbf{x}_t) + \langle \mathbf{x}_* - \mathbf{x}_t, \widehat{\nabla}\mathcal{L}(\mathbf{x}_t)\rangle + \tau_t\|\mathbf{x}_*\|_1 + \frac{1+\gamma}{2}\|\mathbf{x}_* - \mathbf{x}_t\|_2^2 - \frac{1+\gamma}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2^2 \\
&+ \frac{\delta_{3s}^2}{2\gamma}\|\mathbf{x}_t - \mathbf{x}_*\|_2^2 - \tau_t\|\mathbf{x}_{t+1}\|_1 \quad \text{(By optimality of } \mathbf{x}_{t+1} \text{ and the strong conveity)}
\end{aligned}$$

Define

$$\Gamma_t = \frac{\gamma + \delta_{3s}^2/\gamma}{2}\|\mathbf{x}_* - \mathbf{x}_t\|_2^2 + \tau_t\left(\|\mathbf{x}_*\|_1 - \|\mathbf{x}_{t+1}\|_1\right) + \langle \mathbf{x}_* - \mathbf{x}_t, \widehat{\nabla}\mathcal{L}(\mathbf{x}_t) - \nabla\mathcal{L}(\mathbf{x}_t)\rangle - \frac{1+\gamma}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2^2$$

We have

$$\mathcal{L}(\mathbf{x}_{t+1}) \leq \mathcal{L}(\mathbf{x}_t) + \langle \mathbf{x}_* - \mathbf{x}_t, \nabla\mathcal{L}(\mathbf{x}_t)\rangle + \frac{1}{2}\|\mathbf{x}_* - \mathbf{x}_t\|_2^2 + \Gamma_t = \mathcal{L}(\mathbf{x}_*) + \Gamma_t = \Gamma_t$$

where last equality follows from $\mathcal{L}(\mathbf{x}_*) = 0$. Next, we bounded Γ_t by

$$\begin{aligned}
\Gamma_t = & \frac{\gamma + \delta_{3s}^2/\gamma}{2}\|\mathbf{x}_* - \mathbf{x}_t\|_2^2 + \tau_t\left(\|\mathbf{x}_*\|_1 - \|\mathbf{x}_{t+1}\|_1\right) + \langle \mathbf{x}_* - \mathbf{x}_t, \widehat{\nabla}\mathcal{L}(\mathbf{x}_t) - \nabla\mathcal{L}(\mathbf{x}_t)\rangle \\
& - \frac{1+\gamma}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2^2 \\
\leq & \frac{\gamma + \delta_{3s}^2/\gamma}{2}\Delta_t^2 + \frac{\theta_{s,s} + \delta_s + \gamma}{\sqrt{s}}\Delta_t\sqrt{s}\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2 + \langle \mathbf{x}_* - \mathbf{x}_t, (I - U^\top U)(\mathbf{x}_* - \mathbf{x}_t)\rangle \\
& - \frac{1+\gamma}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2^2 \\
\leq & \left(\frac{\gamma + \delta_{3s}^2/\gamma}{2} + \frac{(\theta_{s,s} + \delta_s + \gamma)^2}{2}\right)\Delta_t^2 + \frac{1}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2^2 + \delta_{2s}\|\mathbf{x}_t - \mathbf{x}_*\|_2^2 - \frac{1+\gamma}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2^2 \\
\leq & \left(\frac{\gamma + \delta_{3s}^2/\gamma}{2} + \frac{(\theta_{s,s} + \delta_s + \gamma)^2}{2} + \delta_{2s}\right)\Delta_t^2 - \frac{\gamma}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2^2
\end{aligned}$$

Since $\mathcal{L}(\mathbf{x}_{t+1}) = \|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2^2/2$, we have

$$\frac{1+\gamma}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2 \leq \left(\frac{\gamma + \delta_{3s}^2/\gamma}{2} + \frac{(\theta_{s,s} + \delta_s + \gamma)^2}{2} + \delta_{2s}\right)\Delta_t^2$$

leading to

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2^2 \leq \frac{1}{1+\gamma}\left(\gamma + \frac{\delta_{3s}^2}{\gamma} + 2\delta_{2s} + (\theta_{s,s} + \delta_s + \gamma)^2\right)\Delta_t^2$$

Since δ_s is no-decreasing in s, if we assume $\gamma \geq \max(\delta_{3s}, \theta_{s,s} + \delta_s)$, we have

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2^2 \leq \frac{4\gamma + 4\gamma^2}{1+\gamma}\Delta_t^2 \leq 4\gamma\Delta_t^2 = \Delta_{t+1}^2$$

■

References

Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *CoRR*, abs/1104.4824, 2011.

Kristian Bredies and Dirk A Lorenz. Linear convergence of iterative soft-thresholding. *Journal of Fourier Analysis and Applications*, 14(5-6):813–837, 2008.

R. Garg and R. Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property, 2009.

E. T. Hale, Y. Wotao, and Y. Zhang. Fixed-point continuation for l1-minimization: methodology and convergence. *SIAM J. on Optimization*, 19(3):1107–1130, 2008.

J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

Z. Wen, W. Yin, D. Goldfarb, and Y. Zhang. A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization and continuation. *SIAM Journal on Scientific Computing*, 32(4):1832–1857, 2010.

S. J. Wright, R. D. Nowad, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

Lin Xiao and Tong Zhang. A proximal-gradient homotopy method for the sparse least-squares problem. *CoRR*, abs/1203.3002, 2012.

L. Zhang, M. Mahdavi, R. Jin, and T. Yang. Recovering optimal solution by dual random projection. *ArXiv e-prints*, 2012.